DOCUMENT RESUME

ED 235 246 TM 830 679

AUTHOR Kulick, Edward
TITLE Presentation and Evaluation of a Regression Based Item Bias Detection Technique.
PUB DATE Apr 83
NOTE 24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, April 12-14, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; Cost Effectiveness; Data Analysis; Difficulty Level; *Regression (Statistics); *Research Methodology; *Test Bias; Test Items
IDENTIFIERS Biserial Correlation

ABSTRACT

This paper presents a promising new item bias detection technique that derives practical appeal from its simplicity and economy, and provides data relating to the method's consistency across samples. The proposed regression on item characteristics (RIC) method is a straightforward and inexpensive approach to the study of item bias, based on transformed item difficulties and biserial correlations. It seems particularly well suited for situations of limited resources. It has certain methodological advantages over several competitors apt to be used in similar situations. The RIC method uses multiple regression to predict differences in group performance. Items identified by the proposed method as possibly biased are those items exhibiting group differences which cannot be accounted for by item difficulty, item discrimination, or overall group differences in ability. Results from this study suggest that the RIC method is a promising tool for item bias research although its consistency on small samples (N=125) has yet to be demonstrated. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED235246

Presentation and Evaluation of a Regression Based Item Bias Detection Technique[1,2]

Edward Kulick
Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
E. Kulick
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 830 679

[1] A paper presented at the annual meeting of NCME in Montreal, April 1983.

[2] The author would like to thank Neil J. Dorans for reviewing this paper.

Introduction

Much of the test item bias research conducted to date consists of comparative studies seeking evidence of convergent validity among various procedures. Unfortunately, such evidence has been difficult to come by. One of the few generalizations possible is that several of the most popular methods (e.g., chi-square, delta-plot, and ICC) seem to display moderate levels of agreement (see Ironson and Subkoviak (1979), Shepard, Camilli, and Averill (1980)). It may be the case that in a given situation many item bias procedures are sufficiently adequate to identify a set of the most potentially biased items in a pool. The important question then becomes not Which is best? but, Which is least expensive? or, Which is simplest to use or explain? A procedure which is found to be economical, conceptually simple, and theoretically sound would have obvious value. For example, item response theory methods, though elegant, are often prohibitively expensive. The delta-plot method, while inexpensive, has serious methodological drawbacks.

The purposes of this paper are twofold. First it seeks to present a promising new item bias detection technique, one that derives practical appeal from its simplicity and economy. Secondly, it provides data relating to the method's consistency across samples. Consistency is an important characteristic of any item bias procedure if one is to put faith in the results of a single application. It will be difficult to compare the consistency of the proposed method to other methods since very little research has been done in this area, generally because of the difficulty in obtaining sufficient samples. Many factors affect the consistency of results, including test length, level of bias in the test, and size of samples. Only the last

factor will be varied in the current study. While it is recognized that consistency is a necessary, but not a sufficient condition for a valid item bias procedure, it is not the intent of this paper to study other aspects of the proposed method or to compare it with other existing methods.

Description of the Method

The proposed method presented in this paper was suggested by Richard B. Darlington (personal communication). It will hereafter be referred to as the regression on item characteristics (RIC) method. Other versions of it have been used by Boyce and Darlington (1981) and Kulick (1982). The RIC method is a straightforward and inexpensive approach to the study of item bias, based on transformed item difficulties. The RIC method identifies items as biased if they appear more difficult for one group than another, compared to other items of equal discrimination. Thus the RIC method is a relative technique and its evaluation of a particular item is dependent on the other items included in the analysis. It should be noted that the term bias is used in this paper only to refer to instances of unexpected differential item performance. Items labeled as biased are merely those that deserve closer examination on an individual basis.

Suppose that two groups are selected for a study. Let A and B be defined as the percent of each group responding correctly to an item, transformed using the arcsine transformation[1]. The main purpose of this transformation is to help linearize the relationship between group difficulty values. One

---

[1] If p represents an item p-value, then $A = 2\arcsin(\sqrt{p})$ represents the arcsine transformation used in this study.

measure of group difference in item performance is simply the value of A-B. If groups 1 and 2 perform equally well on an item, Then A-B = 0. Many factors other than bias, of course, may contribute to nonzero values of A-B. The most obvious one is probably a difference in group abilities with respect to the construct being measured. Other factors which can contribute to this difference are item characteristics such as difficulty and discrimination. For instance, the relationship of A to B, despite the arcsine transformation, is not perfectly linear. Thus, differences between A and B are partly a function of the non-linear relationship between group difficulty indices. In addition, if the groups differ in the ability being measured, the discriminating power of the item will affect the value of A-B. Thus, one would not want to use A-B as an index of item bias.

The RIC method uses multiple regression to predict differences in group performance from the factors mentioned above. The RIC method assumes that, in the absence of bias, any two items of equal discrimination and difficulty ought to exhibit the same between-group difference in performance. To the extent that the observed group difference in item performance (A-B), differs from the predicted value ($\widehat{A-B}$), an item is considered biased. In other words, those items identified by the proposed method exhibit group differences which cannot be accounted for by item difficulty, item discrimination, or overall group differences in ability.

The mechanics of the RIC method are simple to perform and easy to explain. First the arcsine transformation is applied to the p-values (percent correct) for each group. Then the biserial correlations (r_a, r_b) are computed

within each group. The biserial correlation serves as a measure of item discrimination. Next A-B is regressed on A+B, $(A+B)^2$, and $(r_a+r_b)/2$. Residuals from the regression serve as indices of each item's bias or unexpected differential item performance. The larger the residual, the greater the level of bias exhibited by that item. By using these predictors the bias index will be uncorrelated with item difficulty (as measured by A+B and $(A+B)^2$) and item discrimination (as measured by $(r_a+r_b)/2$).

Geometrically, one may think of each item as a point in space determined by its values on the A-B, A+B, and $(r_a+r_b)/2$ axes. Imagine trying to find that plane which best fits or summarizes these points (in the least squares sense). (See Figure 1). Since the term $(A+B)^2$ is included in the regression, the plane may bend along the A+B axis in order to more closely approximate the points (items) and account for any non-linearity between difficulty values. Each point's vertical distance from the plane is defined as the index of bias. These residuals reflect the size of A-B after controlling for item difficulty, item discrimination, and group differences in ability.

The arcsine transformation allows for easy significance testing of the indices, should one choose to do so. However, the usefulness of such testing is questionable as it lends pseudo-objectivity to what, at least in part, should be a subjective process. Therefore significance tests are not reported on the indices in this study.

The RIC method appears to have a number of advantages associated with it. First it is conceptually simple. This is important not only to those who use it, but to those who must explain it or have it explained to them. It is also inexpensive. The output from typical item analysis programs provide the input

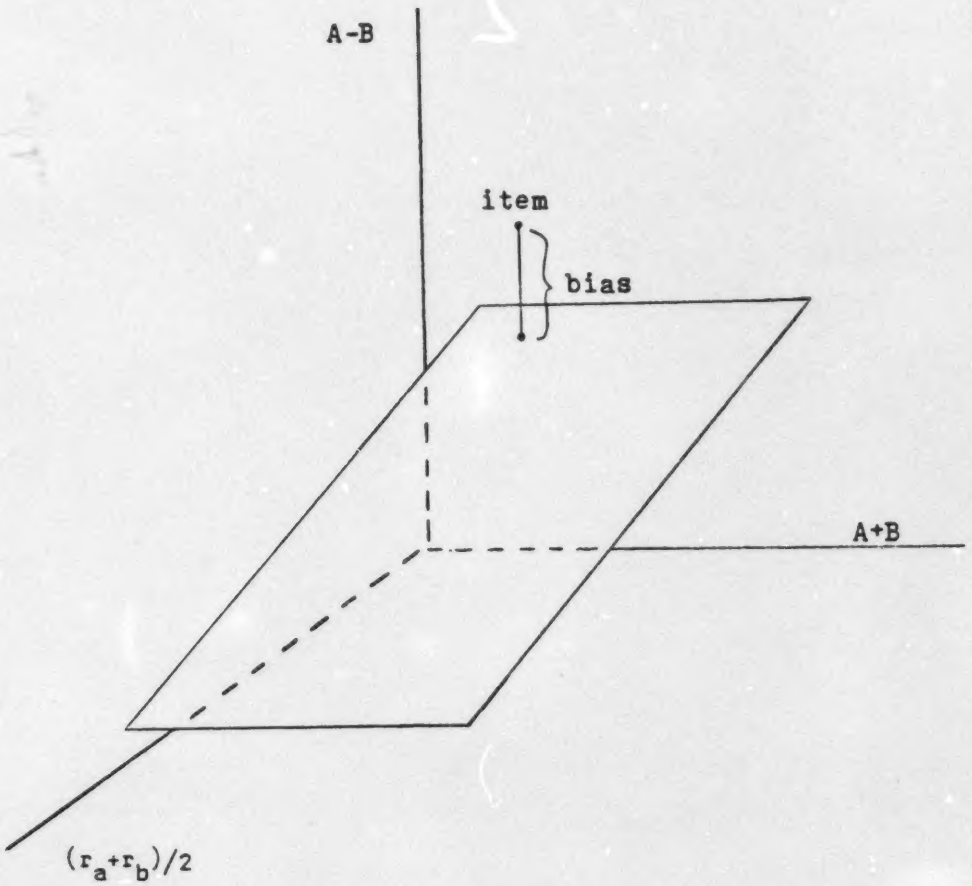

FIGURE 1: Geometric representation of bias index as residual from regression plane.

for the regression analysis. No special or complex programs are required, as is the case for most item response theory techniques. In addition to being inexpensive to perform, regression procedures are available on nearly all statistical packages. Thus accessibility is not a concern. The method should also be applicable in situations where small sample size prohibits the use of some alternative methods. Accurate estimates of p-values can be obtained from samples perhaps as small as 50. The limiting consideration is how large a sample is required for stable estimates of the biserial correlation. In some ways the RIC method is similar to the delta-plot method discussed by Angoff (1982). There is a major difference however. The RIC method attempts to control for the discriminating power of the items, thus preventing highly discriminating items from inaccurately being identified as biased. In a previous study, Kulick (1982) found that removing the biserial correlation from the RIC method produced nearly identical results to the delta-plot method. Thus the RIC method is flexible, and a modification of it may be used if no biserial correlations are available.

The procedure, of course, is not without weakness. It cannot detect instances of non-uniform bias, that is, items that favor different groups at different points along the ability continuum. Also the calculation of the regression line is contaminated by any biased items that are in the pool. In addition, since the RIC method is a relative techique, results are always dependent upon the set of items studied.

Description of the Study

Test Instrument The verbal section of the Scholastic Aptitude Test (SAT) form ZSA5, administered in December 1977, is the source of the items used in this study. The psychometric properties of SAT form ZSA5 are described fully in the test analysis report by Cook and Nutkowitz (1979). Hence, the 85 item verbal test is described only briefly in this paper. The verbal test consists of four distinct item types: reading comprehension, analogies, antonyms, and sentence completion. The estimated reliability of the verbal test was fairly typical at .914, with a scaled score standard error of measurement of 32. The mean equated delta for the verbal section was 11.3, indicating that the test was slightly easier than intended.

Samples This study compares the performance of male and female candidates on the SAT form ZSA5. From the total group of 166,311 to whom the test was offered, a representative random sample consisting of 21,835 males and 21,209 females served as the candidate population for this study. From the group of 21,835 males, 12 random spaced, potentially overlapping, samples were drawn. Three samples were selected at each of the following sample sizes: 1,000, 500, 250, and 125. Note that a candidate may be in more than one of these 12 samples. Similarly, 12 female samples of corresponding sizes were drawn. Each male sample was arbitrarily paired with one female sample of identical size. The RIC procedure was then applied to each of the 12 pairs of samples and the results of these 12 analyses are presented in the next section.

Results and Discussion

There are many ways one could chose to evaluate the consistency of an item bias procedure. Typically, the primary goal of a procedure is to determine the level of bias exhibited by each item, permitting the ranking of items from most favorable (biased for) to least favorable (biased against) with respect to a certain group. One way to evaluate the consistency with which this goal is reached is to compare these item rankings across samples. Alternatively, one could compare subsets of items judged most biased (against a particular group) across samples or against a set of items externally determined to be biased. Still another way would be to compare just those items displaying statistically significant bias.

This last means of evaluation is perhaps the least helpful of those mentioned. Several factors combine to make significance testing in item bias research undesirable. Significance testing is a function of sample size. Thus, with large enough samples, any difference in group performance can be made significant. Furthermore, the notion of significance testing with respect to item bias is misleading. It implies that bias is dichotomous in nature, when in fact it properly belongs on a continuum. Even in the case where no items are identified as significantly biased an investigator may still want to examine those items that are 'most biased.' Significance tests are not reported in this paper since they are of questionable value.

The data reported in this study consist of item rankings across samples. Consistency is appraised through correlations between item rankings and percentages of agreement between sets of the most biased items. Ideally, one

would prefer to compare the results of each analysis to a criterion ranking or a criterion set of biased items. Unfortunately, no such criteria were available for this investigation. Therefore, no absolute conclusions can be reached regarding the consistency of the RIC procedure. Only statements of its relative consistency can be made, because the consistency level of a biased item detection method is dependent on the level of bias in the test, which for the present study is unknown. The less pervasive the bias is in the test, the greater the proportional influence of error on the rankings, particularly among those items near the center of the rank ordering (i.e., those items least biased against either group). Assuming a procedure can detect bias reasonably well, those items with actual bias are apt to appear consistently at the extremes of the rankings, while those items whose bias indices are attributable largely to error will tend to fall near the center of the ranking, in no consistent order. Reliable item bias procedures will exhibit this tendency. Since decisions will generally be made about extreme items only, an item bias procedure should, if nothing else, identify those sets of relatively most biased items consistently. The less bias in the test, however, the more difficult this objective becomes. Therefore, it becomes difficult to appraise the consistency of a method when the test's level of actual bias is unknown.

The results of this study will be presented under two headings. First, results based on comparisons between same-sized samples are given. These are followed by comparisons across samples of different sizes.

Equal Sample Size Comparisons Each analysis performed by the RIC method produced a distinct ranking of items from most biased against females (#1), to most biased against males (#85). For the discussions to follow, let N represent the number of candidates in each group. Three correlations were computed at each sample size. The correlations were highest, as expected, for N=1000, ranging from .79-.85. When the sample size used was cut in half to 500, the correlations ranged from .65-.68. With N=250, the correlations were between .47 and .59. With the number of males and females at only 125 each, the correlations were between .39-.42 (see Table 1). Note, every correlation in the table is significant.

Sets of the 10% and 20% most biased and least biased items against females were defined. Note that the least biased items against females are not those most free from bias, but those most biased against males. This is not to recommend that item bias procedures be used to determine fixed cutoffs such as 10% or 20% of the items. Rather, these cutoffs are only chosen to facilitate comparisons and summarization of data. In practice, empirical cutoffs should be selected based on examination of the distributions or existing needs. The 20% (17 items) and 10% (9 items) most biased items against females as measured by the RIC method for each of the 12 analyses are listed in Table 2. It is noteworthy that item #63 was identified as the most biased item against females in all 9 analyses based on N's of 250 or more, and in one of the three analyses from N=125. The percentages of agreement across runs were computed and recorded in Table 3. These data reveal an impressive level of agreement between analyses. For instance, samples 1 and 2 (N=1,000) agree on 14 of the

TABLE 1

Correlations Between Item Rankings and Between Bias Indices Across Analyses [1]

		N=1,000			N=500			N=250			N=125		
Analysis:		1	2	3	4	5	6	7	8	9	10	11	12
	1	1	.79	.85	.73	.77	.74	.47	.76	.65	.50	.57	.67
N=1,000	2	.81	1	.83	.72	.70	.69	.58	.70	.69	.57	.62	.56
	3	.86	.85	1	.73	.74	.79	.58	.75	.70	.50	.65	.65
	4	.78	.76	.74	1	.68	.68	.37	.63	.62	.48	.63	.51
N=500	5	.79	.75	.76	.74	1	.65	.51	.69	.61	.46	.52	.58
	6	.76	.74	.81	.71	.69	1	.53	.69	.57	.46	.54	.54
	7	.58	.65	.68	.46	.57	.62	1	.47	.49	.35	.45	.32
N=250	8	.73	.77	.78	.65	.70	.71	.57	1	.59	.39	.50	.65
	9	.73	.75	.74	.69	.65	.65	.57	.65	1	.37	.56	.56
	10	.52	.55	.49	.50	.47	.48	.38	.37	.43	1	.39	.42
N=125	11	.62	.65	.66	.67	.59	.57	.46	.49	.60	.45	1	.40
	12	.65	.57	.64	.53	.54	.57	.38	.60	.58	.46	.40	1

[1] Correlations above the diagonal are between item rankings while correlations below the diagonal are between raw bias indices.

TABLE 2

The Twenty Percent (17 items) Most Biased Items Against Females in Each Analysis

	N:	1,000			500			250			125			
	Analysis:	1	2	3	4	5	6	7	8	9	10[a]	11[b]	12	
	1	63	63	63	63	63	63	63	63	63	83	20	63	
	2	16	36	48	85	84	36	29	44	36	63	36	36	
	3	48	44	44	36	36	6	48	43	48	85	63	48	
B	4	65	20	20	20	45	43	54	20	56	65	45	43	
i	5	20	16	29	16	20	29	44	36	16	68	7	8	
a	6	17	48	7	82	82	83	83	56	84	41	55	56	
s	7	7	54	54	48	43	48	41	48	83	82	29	85	
	8	29	56	36	84	29	20	24	16	54	36	38	75	
R	9	36	83	45	65	7	82	17	54	6	78	6	40	(10% most biased)
a														
n	10	83	84	43	54	48	44	62	62	29	28	24	77	
k	11	56	29	82	56	44	65	64	23	44	51	2	7	
i	12	27	45	65	67	17	16	7	38	17	8	16	29	
n	13	45	38	56	29	62	64	23	7	20	62	84	41	
g	14	84	17	16	75	54	54	84	75	4	30	65	45	
	15	44	65	17	6	16	78	55	32	82	6	21	6	
	16	43	81	53	23	56	23	51	45	65	80	21	6	
	17	82	7	6	45	4	7	20	8	7	16	85	84	(20% most biased)

[a]Items 14, 26, and 45 were not included in analysis 10 because the percent correct in the female group for those items was less than .05.

[b]Item 26 was not included in analysis 11 because the percent correct in the female group for that item was less than .05.

TABLE 3

Percentage of Agreement Between Analyses for the 10% and 20% Most Biased Items Against Females

		N=1,000			N=500			N=250			N=125		
Analysis:		1	2	3	4	5	6	7	8	9	10	11	12
	1	-	.55	.66	.55	.55	.55	.44	.55	.44	.33	.55	.33
N=1,000	2	.82	-	.66	.55	.33	.55	.55	.88	.77	.33	.33	.44
	3	.82	.76	-	.44	.55	.55	.55	.66	.44	.22	.66	.33
	4	.65	.65	.71	-	.55	.55	.22	.55	.55	.55	.33	.44
N=500	5	.82	.76	.82	.65	-	.66	.22	.44	.33	.33	.66	.33
	6	.71	.65	.76	.65	.65	-	.44	.55	.55	.44	.55	.44
	7	.53	.59	.47	.41	.59	.59	-	.44	.55	.33	.22	.22
N=250	8	.59	.65	.65	.59	.71	.59	[illegible]		.55	.22	.33	.44
	9	.82	.82	.82	.71	.76	.76	.59	.53	-	.33	.33	.44
	10	.35	.35	.41	.41	.29	.47	.29	.35	.41	-	.22	.33
N=125	11	.59	.65	.59	.65	.53	.53	.47	.47	.59	.35	-	.22
	12	.59	.53	.59	.65	.53	.47	.35	.53	.53	.53	.59	-

Entries above the diagonal are based on the 10% most biased, and entries below the diagonal are based on the 20% most biased against females.

17 most biased items (82%) against females. The percentages of agreement between analyses based on samples of N=1,000 range from .76 to .82. The percent of agreement that would be expected by chance alone is only .20. Analyses based on samples of N=500 all displayed 65% agreement. The range of agreement between runs with N=250 was .41-.59. Finally, runs based on N=125 showed levels of agreement of .35, .53 and .59. In short, the RIC method does extremely well (very consistent) at N=1,000 and also showed fairly high agreement at N=500. Runs based on the smallest two sample sizes showed more variability in the items placed in the top 20%, but were still well above the chance level of agreement.

Inspection of the sets of 10% most biased items versus females showed equally impressive agreement. The chance level of agreement here is only 10%, yet the observed agreement is much greater. Analyses based on N=1,000 showed agreement percentages between 55% and 66%. The agreement between runs based on N=500 ranged from .55-.66; for N=250 from .41-.59; and for N=125 from .22-.33. All runs of the RIC method exhibited greater agreement among the set of most biased items against females than chance levels. Indeed, at some levels, agreement was fairly extensive.

Similarly, sets of the 20% and 10% most biased items against males were determined and are presented in Table 4. The percentages of agreement between runs of the RIC method were calculated and are recorded in Table 5. These percentages are not as high as the corresponding percentages from the classifications of items biased against women, although they are still above chance levels. This difference in percentages of agreement may be a function

TABLE 4

The Twenty Percent (17 items) Most Biased
Items Against Males in Each Analysis

	N:	1,000			500			250			125			
	Analysis:	1	2	3	4	5	6	7	8	9	10[a]	11[b]	12	
	1	69	79	52	52	77	15	10	9	33	49	59	42	
	2	76	1	9	77	1	11	1	1	47	15	15	9	
	3	52	9	47	37	15	3	76	33	9	57	79	5	
B	4	37	77	79	72	10	74	47	13	1	25	72	46	
i	5	1	14	73	80	37	47	50	10	37	42	49	50	
a	6	42	37	3	69	59	1	38	5	13	5	52	57	
s	7	79	33	57	13	13	9	30	31	46	52	9	70	
	8	77	3	1	57	52	71	52	80	72	23	77	55	
R	9	33	52	46	1	73	57	5	69	66	79	8	47	(10% most biased)
a														
n	10	73	27	30	66	60	60	71	79	28	71	47	12	
k	11	80	5	42	15	12	34	57	12	5	44	68	31	
i	12	9	73	5	33	49	10	33	49	80	9	40	79	
n	13	57	57	13	70	5	52	32	25	49	70	46	24	
g	14	3	49	12	79	69	2	49	46	60	72	1	59	
	15	10	74	80	10	42	31	73	81	25	1	12	37	
	16	5	25	15	47	50	69	82	52	59	35	80	68	
	17	12	72	71	49	68	22	18	22	31	33	33	22	(20% most biased)

[a] Items 14, 26, and 45 are not included in analysis 10 because the percent correct in the female group for those items was less than .05.

[b] Item 26 was not included in analysis 11 because the percent correct for the female group on that item was less than .05.

TABLE 5

Percentage of Agreement Between Analyses for the 10% and 20% Most Biased Items Against Males

		N=1,000			N=500			N=250			N=125		
Analysis:		1	2	3	4	5	6	7	8	9	10	11	12
	1	-	.66	.33	.55	.44	.11	.33	.33	.33	.33	.33	.11
N=1,000	2	.65	-	.55	.55	.44	.33	.22	.33	.44	.22	.44	.11
	3	.65	.47	-	.33	.33	.55	.33	.22	.44	.33	.33	.44
	4	.59	.53	.47	-	.55	.22	.22	.44	.44	.22	.33	.11
N=500	5	.59	.41	.47	.53	-	.22	.33	.33	.33	.22	.44	.00
	6	.41	.35	.47	.41	.35	-	.22	.22	.33	.22	.22	.33
	7	.47	.41	.47	.41	.35	.35	-	.44	.22	.22	.11	.22
N=250	8	.59	.47	.53	.53	.47	.41	.35	-	.44	.11	.11	.22
	9	.35	.47	.41	.53	.41	.29	.29	.59	-	.00	.22	.33
	10	.47	.59	.53	.53	.35	.35	.41	.47	.41	-	.44	.33
N=125	11	.47	.47	.53	.59	.47	.29	.29	.53	.53	.47	-	.11
	12	.41	.29	.47	.29	.41	.29	.18	.41	.41	.35	.41	-

Entries above the diagonal are based on the 10% most biased, and entries below the diagonal on the 20% most biased against males.

of the level and direction of the bias present in the test. If little or no bias is present, then item rankings are determined largely by chance. There is no reason to expect that any bias in the test is equally divided between males and females. It may be that the bias against females was more extensive in terms of magnitude and number of items. If so, then the inclusion of items in the 10% and 20% most biased against males categories, is more influenced by chance than for the 10% and 20% against females. Since agreement among the item sets biased against males is above chance levels, the RIC method is nonetheless demonstrating some consistency in measuring the existing level of bias against males. There is reason to suspect that this consistency would increase if the level of bias against males approached the level against females.

Comparisons Across Sample Sizes Thus far comparisons of data have been limited to those based on equal samples. In that way, the consistency of the RIC procedure was appraised at four different sample sizes. The present section examines how closely results from one sample agree with those from different sized samples. Since three item rankings were obtained at each sample size, nine correlations are possible between any two sample sizes. These correlations are presented in Table 1. The correlations between samples of 1,000 and samples of 500 range from .69 to .79, with a median value of .73. It appears that using samples only half as large (500) still produce very similar rank orderings. The correlations between samples of 1,000 and 250 range from .47 to .76, with a median of .69. The correlations between samples

of only 125 and 1,000 range from .50 to .67. with a median of .57. All of these correlations appear fairly strong (and all are statistically significant).

Agreement in determination of most biased items vs. females is also high between different sample sizes. For the 20% cutoff, the ranges in percentage of agreement between samples of 500 and 1,000 were .65 - .82, with a median value of .71. The corresponding ranges and medians for samples of N=250 and N=125 compared to N=1,000 were .47 - .82, .65 and .35 - .65, .59, respectively (see Table 3). For the 10% cutoff the samples of N=250 showed better agreement with samples of N=1,000 than the N=500 samples. The percentages of agreement between N = 1,000 and N = 250 samples (Table 3) range from .44 to .88, with a median of .55. Recall the chance value of these percents is only .10. The range of agreement between samples of 500 and 1,000 is .33 - .55, with .55 also being the median. Agreement between samples of 125 and 1,000 goes from .22 to .66, with the median equal to .33.

Inspection of the 10% and 20% most biased items against males revealed results similar to previous ones. While the percent of agreement between different sample sizes was above the chance levels, it was lower than the agreement exhibited in the sets of items biased vs. females.

Summary and Conclusions

The proposed regression on item characteristics (RIC) method appears to be a promising tool for item bias research. It seems particularly well suited for situations of limited resources. It is conceptually and computationally

simple, economical, and accessible to nearly all potential users. Furthermore, it has certain methodological advantages over several competitors apt to be used in similar situations. Of course no method, including the RIC method, can be recommended for use as the sole determinant of bias in test items. That determination is inherently, at least partly, judgmental. One only hopes that statistical procedures can supply useful data to aid in this process. It seems that the RIC method is capable of that function. Again, throughout this paper bias is used only to mean that an item seems to be functioning in somewhat different ways for different groups.

The data from this investigation are informative, yet conclusions are difficult to reach because no standards have been established with respect to consistency. Little is known about the consistency of competing methods and generalizations concerning how the RIC method would do on other tests are impossible to make. Presumably, the level of consistency would increase directly with the level of bias. In fact, the greater consistency shown by the RIC method involving items biased against females compared to items biased against males may have been a function of the relative magnitude of bias against each group in the test. Consistency was demonstrated at each sample size through correlations of rankings and percentages of agreement in sets of most biased items. Particularly encouraging was the high agreement between sets of items most biased against females across samples of the same size. Predictably, the method performed most consistently on the largest sample, and decreased with each successivley smaller sample. Results from samples of 125 and 250 displayed considerable agreement (greater than chance) with those from

samples of 500 and 1,000. Unfortunately, since no 'true' criterion available it cannot be determined how well the samples did as substitutes for the population. Correlations and percent agreement would presumably be higher with 'true' rankings since sample rankings suffer to an unknown extent from unreliability. Another factor which makes the consistency of the method difficult to appraise is test length. Perhaps the method would show greater consistency on a shorter test. In sum, use of the method on samples of 125 is difficult to recommend based on the correlations of rankings with samples of 1,000, with other samples of 125, the percentage agreement with samples of 1,000 and the failure of two of the three samples of 125 to identify item #63 as most biased against females. This may, however, be more of a failure to demonstrate consistency in this particular study on this particular test, than it is a finding that samples of 125 will always be inappropriate.

More data are needed to compare the consistency of the proposed method to other methods, or to reach conclusions on what acceptable levels should be. Establishing a criterion ranking or set of biased items would enable more conclusive statements to be drawn concerning the consistency of the RIC method. The indefinite nature of the data presented does not diminish the apparent potential of the RIC method to become a valuable procedure for detecting test item bias.

REFERENCES

Angoff, W. H. Use of difficulty and discrimination indices for detecting item bias, chapter 4 in Handbook of methods for detecting item bias (R. A. Berk, Ed.). Baltimore: Johns Hopkins University Press, 1982.

Boyce, C., & Darlington, R. B. Black proficiency in abstract reasoning. Unpublished manuscript, 1981. (Available from authors, Psychology Dept., Uris Hall, Cornell University, Ithaca, NY 14853.)

Burrill, L. E. Comparative studies of Item Bias Methods, chapter 6 in Handbook of methods for detecting item bias (R. A. Berk, Ed.). Baltimore: Johns Hopkins University Press, 1982.

Cook, L., & Nutkowitz, I. Test Analysis: College Board Scholastic Aptitude Test December 1977 Administration SZA5 (SR-79-63). Princeton, NJ: Educational Testing Service, 1979.

Devine, P. J. & Ragu, N. S. Extent of overlap among four item bias methods. Educational and Psychological Measurement, 1982, 42, 1049-1066.

Ironson, G. H., & Subkoviak, M. J. A comparison of several methods of assessing item bias, Journal of Educational Measurement, 1979, 16, 209-255.

Kulick, E. Evaluation of a regression based method for studying test item bias. Unpublished Masters Thesis, Cornell University, 1982.

Rudner, L. M., Getson, P. R., & Knight, D. L. A Monte-Carlo investigation of seven biased item detection techniques. Journal of Educational Measurement, 1980, 17, 1-10.

Scheuneman, J. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Shepard, L., Camilli, G. & Averill, M. Comparison of six procedures for detecting test item bias using both internal and external ability criteria. Paper presented at annual meeting of National Council on Measurement in Education, Boston, April, 1980.

Abstract

"Presentation and Evaluation of a Regression Based
Item Bias Detection Technique"

Edward Kulick

The purposes of this paper are two-fold. First it seeks to present a promising new item bias detection technique that derives practical appeal from its simplicity and economy. Secondly, it provides data relating to the method's consistency across samples.

The proposed regression on item characteristics (RIC) method is a straight forward and inexpensive approach to the study of item bias, based on transformed item difficulties and biserial correlations. It seems particularly well suited for situations of limited resources. It has certain methodological advantages over several competitors apt to be used in similar situations.

The RIC method uses multiple regression to predict differences in group performance. Items identified by the proposed method as possibly biased are those items exhibiting group differences which cannot be accounted for by item difficulty, item discrimination, or overall group differences in ability.

Results from this study suggest that the RIC method is a promising tool for item bias research although its consistency on small samples (N=125) has yet to be demonstrated.